Empirical Feasibility of the Desirable Difficulty Framework: Toward More Systematic Research on L2 Practice for Broader Pedagogical Implications

YUICHI SUZUKI,[1] TATSUYA NAKATA,[2] and ROBERT DEKEYSER[3]

[1]*Kanagawa University, Faculty of Foreign Languages, 3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa, 221-8686, Japan Email: szky819@kanagawa-u.ac.jp*

[2]*Hosei University, Faculty of Letters, 2-17-1 Fujimi, Chiyoda-ku, Tokyo, 102-8160, Japan Email: t-nakata@hosei.ac.jp*

[3]*University of Maryland College Park, School of Languages, Literatures, and Cultures, 3104 Jimenez Hall, College Park, MD, 20742 Email: rdk@umd.edu*

IT WAS OUR PLEASURE TO RECEIVE A commentary from Rogers and Leow (this issue) on Suzuki, Nakata, & DeKeyser (2019), the coda chapter of the special issue (*Optimizing Second Language Practice in the Classroom: Perspectives from Cognitive Psychology*) in *Modern Language Journal 103*(3). Their commentary was overall positive toward a theoretical framework for systematic and deliberate second language (L2) practice that we proposed in the chapter. Our framework is derived from (a) the cognitive difficulty framework in the area of L2 acquisition research (Housen & Simoens, 2016) and (b) the desirable difficulty framework proposed in the realm of cognitive psychology (Bjork, 2018). Rogers and Leow's commentary consists of the following three major components: (a) the lack of descriptions of concurrent learning processes, (b) empirical (in)feasibility of our framework, and (c) difficulty in drawing pedagogical implications from empirical research. In this response article, we will address each of these concerns in order.

EXPOUNDING ON LEARNING PROCESSES DURING L2 PRACTICE

In their commentary, Rogers and Leow emphasized the importance of examining concurrent learning processes, while referring to Table 1 in Suzuki et al. (2019, p. 716). In the table, some conditions (e.g., blocked practice, shorter spacing, retrieval) are labeled as inducing "low difficulty," while others (e.g., interleaved practice, longer spacing, trial and error) are associated with "high difficulty." Rogers and Leow argued that "Although intuitive, such a categorization may be problematic given a lack of empirical evidence to validate the relative difficulty of the different learning conditions" (p. XX).

We believe that their criticism does not apply because it is possible to estimate the relative difficulty of different learning conditions based on empirical evidence—namely, learning phase performance. In our coda chapter, we pointed out that "Accuracy/error data *during* practice (e.g., proportion of correct responses on grammar exercises [Nakata & Suzuki, in this issue] or during vocabulary exercises [Strong & Boers, in this issue]) can be used as a measure of L2 difficulty" (Suzuki et al., 2019, p. 714). For instance, Nakata and Suzuki (2019b) demonstrated that interleaved practice led to a significantly lower proportion of correct responses during the learning phase (77.0%) than blocked practice (87.2%). Similarly, Strong and Boers (2019) reported that

the retrieval condition led to more successful performance during learning (88%) than the trial-and-error condition (18%). Kasprowicz, Marsden, and Sephton (2019) did not find a statistically significant difference between long and short spacing in learning phase performance (7-day: 79.6%, 3.5-day: 82.5%). However, a number of studies demonstrate that long spacing induces more errors during the learning phase than short spacing (e.g., Brown, Roediger, & McDaniel, 2014; Nakata & Suzuki, 2019a; Nakata & Webb, 2016; Pashler, Zarow, & Triplett, 2003; Toppino & Gerbier, 2014). By comparing learning-phase performance of different learning conditions, we believe that it is possible to empirically determine their relative difficulty.

Rogers and Leow also raised concern about the use of judgment of learning data in Nakata & Suzuki (2019b). They stated that:

> It should be noted that one study (Nakata & Suzuki, 2019) in the collection attempted to account for difficulty by collecting judgment of learning data from the participants in the study. However, as these data were collected during the testing phase of the study, they fall prey to the criticism of using a retrospective instrument as a means of measuring a concurrent process, thus calling into question the internal validity of the findings here (see Leow & Hama, 2013, for a discussion). (Rogers & Leow, this issue, p. XX)

We completely agree that judgment of learning data cannot be used as an objective, concurrent measure of learning difficulty. However, we never intended to claim that they should be. The purpose of examining judgment of learning data in Nakata & Suzuki (2019b) was to examine retrospective, *perceived* (subjective) difficulty of training on the part of learners, and not to estimate a concurrent, objective measure of learning difficulty. This was stated in our coda chapter as follows:

> In addition to the objectively measured data, perceived ease or difficulty of training on the part of learners may also be a valuable index of difficulty. The perceived difficulty is closely tied to the perceived effectiveness of training; learners tend to consider training to be effective when difficulty during training is low and learning-phase performance is successful (but see Nakata & Suzuki, in this issue, for the "judgement of learning" data), although successful learning-phase performance does not often yield the most successful outcomes (e.g., Bjork, 1994). (Suzuki et al., 2019, p. 714)

Furthermore, Rogers and Leow pointed out that relative difficulty levels experienced by L2 learners may be examined by using techniques such as think-aloud protocols, reaction times, or eye tracking. We completely agree with this statement. In our coda chapter, we stated the following: "Perhaps, in future research, new technologies such as eye tracking (Révész & Gurzynski–Weiss, 2016) and pupillometry (Schmidtke, 2018) can also provide more sensitive measures of difficulty experienced by learners" (Suzuki et al., 2019, p. 714). Triangulation of these concurrent measurements with accuracy and reaction time data during training may lead to a more comprehensive picture of learning processes.

Last, Rogers and Leow drew on Leow's (2015) depth-of-processing model for highlighting the importance of examining the learning processes, especially in the intake stage. We appreciate their insight and concur that the model is useful for elaborating what constitutes desirable difficulty during the learning phase. Leow's model postulates that depth of processing (e.g., amount of cognitive effort, use of prior knowledge, hypothesis testing, rule formation) and a high level of awareness play key roles in facilitating intake. This perspective indeed holds a good promise for expounding on "different learning processes during L2 practice in quest of complex interactions between multiple variables" (Suzuki et al., 2019, p. 718). Here, by introducing two recent studies on L2 practice, we would like to extrapolate our framework and draw on Leow's depth-of-processing model to illustrate how multiple variables (practice condition, linguistic difficulty, and individual differences) interact to influence desirable difficulty during the practice phase.

Pulido and Dussias (2019) examined the learning of English congruent and incongruent collocations by first-language (L1) Spanish speakers. Congruent collocations refer to expressions that can be literally translated from one language to another. For instance, a Spanish collocation *llevar su nombre* (lit. 'carry his name') corresponds to *carry his name* in English. In contrast, *run a business* is an example of an incongruent collocation because a Spanish collocation *llevar un negocio* (lit. 'carry a business') needs to be translated as *run a business*, not *carry a business*, in English. In the study conducted by Pulido and Dussias, during practice, for incongruent collocations, participants in the L1-interference group were exposed to a distractor that was the literal translation of the L1 collocate (e.g., *carry* for *business*), in addition to the correct answer (e.g., *run* for *business*). For participants in the unrelated group, a distractor was not related to the L1 equivalent (e.g., *touch* for *business*). Pulido and Dussias found that for incongruent collocations, the L1-interference group initially showed slower

reaction times during practice, compared with the unrelated group. On posttests, however, the L1-interference group outperformed the unrelated group for incongruent collocations. Their findings suggest that prompting learners to reflect on their prior knowledge (L1) induced more interference and created a more challenging practice condition for L2 collocation learning (cf. Nakata & Suzuki, 2019a). The learning processes during practice inferred here are interference between L2 and L1, and this can also be explained by the depth-of-processing model (e.g., cognitive effort and prior knowledge use).

Sato and McDonough (2019) is an excellent piece of classroom research on L2 grammar practice, although no explicit reference to the desirable difficulty framework is made. Unlike most empirical studies on L2 grammar practice, which have been carried out in decontextualized laboratory settings (e.g., McManus & Marsden, 2019; Suzuki, 2017; Suzuki & DeKeyser, 2017), Sato and McDonough examined the extent to which prior declarative knowledge about the target structures (*wh*-questions) facilitated proceduralization through contextualized grammar practice in an English-as-a-foreign-language classroom. Learners engaged in a number of teacher-led interactive communicative activities eliciting *wh*-questions that lasted over 5 weeks. The participants showed overall improvement in accuracy and fluency (as measured by speech rate and pauses) after the communicative practice. Of interest to the desirable difficulty framework, all learners in their study had declarative knowledge about the target structures, but to varying degrees, prior to the treatment. The results showed that although higher declarative knowledge resulted in more accurate and fluent production of the target structures during the practice phase, declarative knowledge failed to predict the final outcome of practice.

Skill acquisition theory, according to which declarative knowledge facilitates knowledge proceduralization, can account for learning processes during the practice phase. That is, learners with higher declarative knowledge were able to engage in the target behavior (using the target *wh*-questions accurately) during the initial phase of communicative grammar practice. From the desirable difficulty perspective, however, the explanation of the final outcome may be twofold. On the one hand, learners who already had higher declarative knowledge at the outset obviously still benefited from this after the practice. On the other hand, it would also mean that they might have experienced less difficulty (hence, processed less deeply) during practice, and therefore benefited less from the less challenging practice. To that end, the positive and negative effects of having started with more declarative knowledge might have cancelled each other out, which could have resulted in no significant effect of declarative knowledge on the final outcome.

In our view, these two recent studies help address Rogers and Leow's concern, because they shed light on "the cognitive processes triggered by different practice conditions" (Rogers & Leow, this issue, p. XX) in Suzuki et al.'s (2019) framework, one in the laboratory (Pulido & Dussias, 2019) and the other in classroom settings (Sato & McDonough, 2019).

EMPIRICAL FEASIBILITY: COLLECTING SNAPSHOTS OF COMPLEX PICTURES OF L2 PRACTICE

Rogers and Leow also raised concerns over the empirical feasibility of the framework proposed by Suzuki et al. (2019). As the title of their commentary ("Towards Greater Empirical Feasibility of the Theoretical Framework for Systematic and Deliberate L2 Practice") suggests, they seem to consider this feasibility issue to be the most critical one. Regarding this point, Rogers and Leow stated the following:

> The authors noted that it is "nearly *infeasible* [emphasis added] to define the absolute difficulty level that yields optimal practice condition" (Suzuki et al., 2019, p. 715). In this sense, it may likely be impossible to identify the optimum practice condition. However, the important question is whether it is possible to validly identify the relatively superior practice condition under such a framework. (p. XX)

Unfortunately, our statement seems to be taken out of context because the second half (italicized) was omitted. The original passage read as follows:

> It is, however, nearly infeasible to define the absolute difficulty level that yields optimal practice condition, *as L2 difficulty is best captured in relative terms.* The crux of the current proposal illustrated in Figure 1, then, is that all three major difficulty-related factors (i.e., linguistic, learner-related, and context-related) need to be taken into account for creating the optimal levels of L2 practice (tailored systematic and deliberate practice). (Suzuki et al., 2019, p. 715, emphasis added)

Our claim was that whether a particular practice condition (e.g., interleaved practice, long spacing) induces the appropriate level of difficulty (i.e., desirable difficulty) and yields optimal practice condition is affected by a number

of other linguistic or/and learner-related factors, and it may not be always possible to recommend one particular type of practice for any learner, for any linguistic structure. For instance, when the target structure is formally and conceptually simple (linguistic difficulty is low) and the learners have high prior knowledge, aptitude, and/or motivation, a relatively demanding condition (e.g., interleaved practice, long spacing) may induce desirable difficulty and constitute the optimum practice condition. In contrast, when the target structure is formally and conceptually complex (linguistic difficulty is high) and the learners have low prior knowledge, aptitude, and/or motivation, a relatively demanding condition may be undesirably difficult. Instead, a less demanding condition (e.g., blocked practice, short spacing) may induce the appropriate level of difficulty and facilitate learning. Our original intention was that it is fruitful to take multiple factors (linguistic, learner-related, and context-related factors) into account for identifying the optimal practice condition. It may be too pessimistic to draw a hasty conclusion that it is infeasible to identify the optimum practice condition under our framework, as Rogers and Leow suggested.

Furthermore, Rogers and Leow challenged the validity of our framework by referring to inconsistent findings reported for interleaving effects (Nakata & Suzuki, 2019b; Suzuki & Sunada, 2019) and distributed practice effects (Kasprowicz et al., 2019; Li & DeKeyser, 2019). However, we believe that these inconsistent findings actually support, rather than refute, our framework, because these studies demonstrate how linguistic, learner-related, and context-related factors interact with each other to affect L2 acquisition, which is the basic tenet of our framework. Specifically, Nakata and Suzuki (2019b) and Suzuki and Sunada (2019) examined the effects of the increasing schedule (blocked practice followed by interleaved practice) on L2 grammar acquisition. While Suzuki and Sunada found the superiority of the increasing schedule over both blocked and interleaved practice, Nakata and Suzuki failed to do so. Suzuki et al. (2019) argued that these seemingly contradicting findings may be reconciled by taking into account levels of prior knowledge. In other words, as indicated by the average pretest score (69.8%) on the grammaticality judgement task, the participants in Nakata and Suzuki's study had a relatively high level of prior knowledge of the target structures. As a result, the interleaved practice, which is more difficult than the increasing schedule, perhaps provided the appropriate level of difficulty. In contrast, learners in Suzuki

and Sunada's study demonstrated a lower level of prior knowledge as measured by the oral picture-description pretest (27.0%). As a result, the increasing schedule, which is less demanding than the interleaved schedule, might have led to desirable difficulty and turned out to be the optimal condition.

Given the complexity involved in L2 learning in classrooms, it may be too optimistic to assume that it is possible for researchers to recommend one particular type of practice for any learners, for any linguistic structures. In most cases, the answer will likely be "it depends." Of course, this does not mean that researchers should give up on their pursuit of the optimal practice condition. Instead, they should attempt to delve into the center of a myriad of factors intertwined with each other that affect L2 acquisition and strive to identify the condition that is most likely to work best for a particular group of learners, for learning a particular linguistic target, in a particular context. The framework we proposed in Suzuki et al. (2019) represents one initial step toward better understanding this complex phenomenon of L2 learning.

Rogers and Leow also argued that a major issue with our framework is that it cannot be tested empirically because "it seems very unlikely that a single study, even under the strictest laboratory conditions, would be able to control for all of the potential permutations of the variables included in the framework" (p. XX). It was not our intention, however, to prompt researchers to examine *all* factors included in our framework in a single study. Rather, it would be more reasonable and feasible to focus on a few key variables from the framework at one time. For instance, researchers may choose to investigate the learning of the plural morpheme *–s* in English and examine how the practice condition and learner-related factors may affect learning. In this case, researchers are keeping one of the three components in our framework (i.e., linguistic difficulty) constant, and focusing on the other two (i.e., practice condition and learner-related factors). Replicating the study, while systematically manipulating one or more variables in the framework (e.g., using a different linguistic structure to manipulate linguistic difficulty), may potentially lead to a more comprehensive picture regarding how a combination of factors influence L2 acquisition. We believe that the framework proposed in Suzuki et al. (2019) will guide researchers in this endeavor and at the same time provide them with opportunities to empirically test the validity of the framework along the way.

By being comprehensive and perhaps too ambitious in our framework, we also hoped that researchers would be encouraged to examine interactions of multiple factors that might affect L2 acquisition. This, of course, would require a rigorous and often complex research design that allows researchers to collect only snapshots of complex pictures at a time. It does not mean, however, that it is infeasible.[1] Although such an endeavor no doubt requires an extensive amount of systematic effort from a network of researchers (Morgan–Short et al., 2018), we believe that it will eventually lead to a more comprehensive understanding of complex issues related to L2 practice, which in turn results in more practical and nuanced pedagogical recommendations.

BROADER PEDAGOGICAL IMPLICATIONS: CLASSROOM TEACHING AND BEYOND

As acknowledged by Rogers and Leow, the main goal of our coda chapter was not to discuss pedagogical implications derived from the empirical research in the special issue, but to propose an overarching framework of optimizing L2 practice. For pedagogical implications, we refer the readers to Lightbown's (2019) commentary, where she provides an excellent and down-to-earth discussion on how L2 practice research in the special issue can be informative for L2 teachers. At the same time, we would like to clarify the extent to which existing (or future) research on L2 practice can speak to pedagogical implications, because we concur with Rogers and Leow that "pedagogical implications should receive serious consideration as part of any discussion of such L2 practice and/or future ISLA [instructed L2 acquisition] research" (p. XX). Furthermore, researchers are responsible for better communicating research findings to teachers (Marsden & Kasprowicz, 2017; Paran, 2017; Sato & Loewen, 2018, 2019).

Ideally, researchers should be able to make concrete pedagogical recommendations such as: One practice condition (e.g., spaced practice) is always more effective than another (e.g., massed practice). However, given the myriad of factors that may affect L2 acquisition, it is not always possible, or perhaps appropriate, to make such a blanket statement. Instead, researchers should attempt to understand how a number of factors (e.g., linguistic or learner-related) may interact with each other to affect L2 acquisition. A deeper understanding of this complex phenomenon of L2 practice might provide teachers with opportunities to critically reflect on their current teaching methods (Rankin & Becker, 2006) and adjust their potential bias regarding perceived effectiveness of teaching techniques. L2 research has proven to provide useful insights into pedagogy, which is often constrained by individual teachers' beliefs and experiences (Paran, 2017). For instance, although some instructors are hesitant to use L1 translation for vocabulary teaching, empirical research suggests that providing L1 translation can be an effective teaching option (Paran, 2017). Research examining the desirable difficulty framework can be construed as another example. These studies have indicated that a condition that facilitates the initial rate of acquisition (e.g., massing, short spacing, or blocked practice) does not necessarily increase long-term retention (e.g., Nakata, 2015; Nakata & Suzuki, 2019a, 2019b; Pashler et al., 2003). Raising awareness about these somewhat counter-intuitive findings may prompt instructors and learners to readjust their assumptions about L2 teaching and learning. Needless to say, more empirical and theoretical work needs to be done before researchers are able to make more concrete pedagogical recommendations that are useful for instructors and learners. **Q10**

Rogers and Leow were primarily concerned with pedagogical implications "within an instructed setting with curricular goals" (p. XX). We do consider it important to contemplate on how L2 practice can be effectively incorporated into classroom settings. At the same time, potential applications of L2 practice research include not only those for classroom teaching and learning but also for other contexts such as computer-mediated learning (Bower & Rutson–Griffiths, 2016; Lin & Lin, 2019) and materials development (Tomlinson, 2016). As shown in Figure 1 in Suzuki et al. (2019, p. 715), for practice to be truly optimal, it needs to be tailored to individual learners. Tailored practice can perhaps best be implemented using technology-mediated (e.g., computer, mobile device) learning (DeKeyser, 2017). With advances in artificial intelligence, it would not be unrealistic to envisage the potential of tailoring instructions to learners' characteristics. Furthermore, L2 practice research also has the potential to inform material development. For instance, although interleaved practice has been found to be more beneficial for L2 grammar learning than blocked practice (Nakata & Suzuki, 2019b), blocked practice of grammar is common in L2 materials. The findings suggest that material developers may benefit from incorporating interleaved practice into coursebooks. In the end, we emphasize again that seeking pedagogical

conceptual replication and extension. *Language Learning*, *67*, 512–545.

Suzuki, Y., & DeKeyser, R. M. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*, 166–188.

Suzuki, Y., Nakata, T., & DeKeyser, R. M. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *Modern Language Journal*, *103*, 713–720.

Suzuki, Y., & Sunada, M. (2019). Dynamic interplay between practice type and practice schedule in a second language: The potential and limits of skill transfer and practice schedule. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263119000470

Tomlinson, B. (2016). *SLA research and materials development for language learning*. New York: Routledge.

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *The Psychology of Learning & Motivation*, *60*, 113–189.